Journal of Zhejiang University-SCIENCE B (Biomedicine & Biotechnology)   in press
www.jzus.zju.edu.cn; www.springer.com/journal/11585
E-mail: jzus_b@zju.edu.cn

Research Article

https://doi.org/10.1631/jzus.B2400339

Genomic insights into the diversity of rice cultivars developed in Heilongjiang Province of China

Yuhan ZHOU[1], Naixin LIU[2], Jiaqi YANG[2], Baicui CHEN[3], Chengxin LI[3], Fanshan BU[3], Sanling WU[4], Ziqi ZHOU[1], Qingtao YU[3]✉, Qingyao SHU[1]✉

[1] *State Key Laboratory of Rice Biology & Breeding, Zhejiang Provincial Key Laboratory of Crop Germplasm Innovation and Utilization, The Advanced Seed Institute, Zhejiang University, Hangzhou, China*
[2] *Beet Quality Inspection and Test Center Ministry of Agriculture and Rural Affairs, College of Advanced Agriculture and Ecological Environment, Heilongjiang University, Harbin 150080, China*
[3] *Harbin Academy of Agricultural Sciences, Harbin 150080, China*
[4] *Analysis Center of Agrobiology and Environmental Sciences, Faculty of Agriculture, Life and Environment Sciences, Zhejiang University, Hangzhou, China*

Abstract: Amid the fast increase of global population and the quest for sustainable agriculture, the need for enhanced rice breeding strategies has become increasingly pronounced, particularly in Heilongjiang, China's foremost rice-producing province, renowned for its premium temperate japonica rice. Here, we conducted an extensive genomic investigation of the elite rice cultivars developed in Heilongjiang Province. Using whole-genome re-sequencing of a total of 376 representative cultivars from Heilongjiang, of which 14 were developed by a single research group, we identified 4.9 million SNPs and 0.98 million InDels, offering a comprehensive perspective on genetic diversity and population structure. We classified the 376 rice cultivars into five subgroups based on their breeding years. Recently bred cultivars, assigned to subgroups HLJ-IV-1 and HLJ-IV-2, showed notable genetic differentiation. Through a selective sweep analysis, significant genomic variation in genes such as *OsACBP5*, *Os4CL5*, and *GFR1* was pinpointed, reflecting a concerted effort in selecting for broad-spectrum disease resistance and enhanced tillering capacity. Furthermore, to identify the strengths and areas for improvement within those series, we conducted an exhaustive analysis of aromatic compounds and their corresponding genes *OsODC* and *OsBadh2,* as well as the the advantageous long-grain gene *OsGL3.1* haplotype within Hagengdao7. Additionally, strategies for reducing plant height through the introduction of the *sd1* gene were elucidated. With a commitment to expediting the development of superior rice cultivars, our discoveries are poised to raise the sensory attributes and nutritional profile of rice, thereby bolstering the resilience and sustainability of global food systems.

Key words: *Japonica* rice; Heilongjiang; Genomic diversity; Rice breeding

---

✉ Qingyao SHU, qyshu@zju.edu.cn
Qingtao YU, 13694509962@139.com
Yuhan ZHOU, https://orcid.org/0009-0008-3649-4184
Qingtao YU, https://orcid.org/0009-0004-8259-9260
Qingyao SHU, https://orcid.org/0000-0002-9201-0593

Received July 5, 2024; Revision accepted Oct. 21, 2024;
Crosschecked xxx. xx, 20xx;

1 Introduction

In the past half-century, threefold increases in global crop production, notably in rice, have been achieved mainly through genetic improvements via plant breeding, a necessity driven by the fast increase of population and shifting climates (Zheng et al., 2024). This laborious process, traditionally reliant on morphological and phenotypic assessments, spans 8 to 12 years from the selection of parental lines to the commercial release of new cultivars. Recent advancements in molecular markers and genomic selection, fueled by the dramatic cost reductions in next-generation sequencing and the availability of high-quality reference genomes, have revolutionized rice breeding. These technological breakthroughs offer a promising avenue to overcome the intrinsic challenges of traditional breeding methods, facilitating the rapid identification and selection of superior rice genotypes by capturing the subtle genetic nuances critical for enhancing adaptability and yield (Ikeda et al., 2013).

In the context of global rice cultivation, northeast China (NEC), encompassing Heilongjiang (HLJ), Jilin (JL), and Liaoning (LN) provinces, represents an important frontier, particularly HLJ, which alone accounts for about 72% of NEC's rice production and stands as China's preeminent rice-producing province (You et al., 2021). This region, situated at the northernmost viable latitudes for rice cultivation (38.7-53.5°N), has emerged as a crucial zone for the production of high-quality *japonica* rice, responding to consumer demand for superior eating quality (Xin et al., 2020). Despite its relatively recent history of large-scale cultivation, initiated in the 19th century by Korean migrants and further developed through mid-20th century introductions from Japan and Korea, NEC has significantly expanded its rice planting area and production (Chen et al., 2023).

Chen et al. (2023) collected 546 NEC rice cultivars, of which 309 originated from HLJ Province. These cultivars were divided into several subgroups according to different planting areas. This information is invaluable for breeding-by-design, allowing for the targeted selection of desirable traits to develop new rice cultivars with enhanced performance. However, this study lacked a comprehensive analysis of the diverse cultivars present in HLJ. Additionally, the classification of subpopulations for HLJ's cultivars was based mainly on their year of breeding, and the most recent breeding cultivars were not included. This calls for more studies on rice cultivars developed in HLJ and for further exploration of the population structure and genetic basis of selected agronomic traits.

Among the renowned rice cultivars of HLJ Province, Daohuaxiang (DHX) rice, produced mainly from the Wuyoudao 4 cultivar in Wuchang City, is particularly celebrated for its distinctive taste and substantial economic value. Highlighted in a 2021 study, DHX rice has earned widespread acclaim across China for its exceptional quality (Jie et al., 2021). Despite the DHX brand commanding a premium—about 70% higher than standard rice—the province faces a unique dilemma. While it has demonstrated its capability to produce high-caliber rice, enhancing both farmer incomes and company returns, there is a notable lack of similarly esteemed cultivars. This deficiency hampers HLJ's ability to further raise its status as a top rice-producing region and fully meet the demand for premium rice products. This challenge accentuates the need for HLJ to broaden and improve its assortment of elite rice cultivars. Cultivars such as Longgeng31 (Liu et al., 2021) and Kongyu131 (Wang et al., 2023) are also noteworthy for their quality and resilience, offering attributes such as cold tolerance and disease resistance. However, to truly revolutionize the agricultural landscape, the focus must shift towards the development of new cultivars using advanced breeding techniques, including genomic selection and marker-assisted selection.

The Harbin Academy of Agricultural Sciences has developed a range of derivative rice strains with DHX as the core, collectively named the Hagengdao rice cultivar series. As a significant new aromatic rice cultivar series known for its high quality, Hagengdao has undergone numerous generations of selection in the northern regions and extensive multi-regional identification tests over several years. Each new cultivar has demonstrated excellent yield performance and technological maturity. Additionally, these cultivars exhibit superior qualities in terms of appearance and taste, closely resembling the renowned DHX in fragrance, grain shape, and taste. The successful development of the Hagengdao series not only highlights its excellence but also provides valuable insights into the genetic and agronomic traits that contribute to high-quality aromatic rice production.

This study aimed to further investigate the rice cultivars developed in HLJ Province, including the Hagengdao series, to reveal the advancements achieved in rice breeding and agronomy. Here, we report the whole genome re-sequencing of 14 Hagengdao series elite *japonica* rice cultivars. Together with the genomic data of 362 cultivars that were resequenced in previous studies (Chen et al., 2023; Liu et al., 2021; Ye et al., 2022), we decoded the structural and functional genomic variations, with the aim of fostering whole genome sequence-driven breeding in rice and setting a paradigm for other crop species.

2 Materials and methods

2.1 Genomic DNA isolation and genome resequencing

A set of 15 temperate *japonica* accessions were sourced from northeast China, comprising 14 cultivars from the Hagengdao series and one elite line 'Hazhandao' originating from HLJ's first Accumulated Temperature Belts Harbin (45°N). Then, we used whole genome sequencing (WGS) to re-sequence these elite *japonica* rice cultivars. Genomic DNA was extracted from young leaf samples using the Hieff NGS® OnePot Pro DNA Library Prep Kit V2. For the construction of a paired-end sequencing library for each elite cultivar, about 10 μg of genomic DNA was used, adhering to the standard pipeline provided by Illumina. The libraries featured an insert size ranging between 300 and 500 bp, with a read length of 150 bp. Sequencing was conducted on the Illumina® NovaSeq 6000 platform, following the manufacturer's prescribed protocols.

2.2 SNP calling

Quality checks on the raw sequence reads from all rice cultivars were conducted using Trimmomatic (version 0.39, available at https://github.com/usadellab/Trimmomatic) (Bolger et al., 2014), after which, high-quality paired-end reads meeting the QC criteria were aligned against the Nipponbare rice reference sequences (T2T-NIP, AGIS1.0) (Shang et al., 2023) using BWA (version 0.7.17-r1188) (Li and Durbin 2009; Li and Durbin 2010). Picard tools (version 3.1.1) were used to sort the aligned reads and remove duplications. SAMtools (Li et al., 2009) (version 1.9) was applied to process the mapping outcomes. Single nucleotide polymorphism (SNP) calling was performed using the Haplotype Caller feature of the Genome Analysis Toolkit (GATK) (DePristo et al., 2011) (version 4.0.5.1), with initial filtering of SNPs identified by GATK set as follows: DP > 24,000, QD < 2.0, FS > 60.0, MQ < 20.0, MQRankSum < -12.5, and ReadPosRankSum < -8.0. Further removal of low-quality variants was based on: (1) a missing rate exceeding 80%, (2) a heterozygous genotype frequency above 5% or more than double the minor homozygous allele frequency, and (3) deviation from Hardy-Weinberg equilibrium as outlined by GATK (excess heterozygosity $<10^{-5}$) among the 376 elites. SnpEff (Cingolani et al., 2012) (version 5.2) was used for the annotation of SNPs to assess their potential functional effects.

2.3 Population structure analysis

The SNP filtering methods described in the study can be summarized as follows: first, to construct the basic dataset, high-quality bi-allelic SNPs were filtered from the total SNP dataset, followed by the removal of SNPs with heterozygosity levels exceeding Hardy–Weinberg expectations. Subsequently, SNPs with 20% missingness and a MAF less than 1% were removed from the basic SNP set to form the filtered dataset. Finally, using PLINK (Purcell et al., 2007), a two-step linkage disequilibrium filtering was performed: (1) with a window size of 10 kb, a SNP window step of one, and an r^2 threshold of 0.8; (2) with a window size of 50 SNPs, a SNP window step of one, and an r^2 threshold of 0.8 (Wang et al., 2018). The resulting core SNP set contained a total of 56,376 SNPs, which were used as input files for subsequent population structure analysis.

To reveal the possible population structure of the rice cultivars tested, multiple analyses, such as model-based population structure analysis, and neighbor-joining tree clustering method, were used to group the cultivars into sub-populations. Population structure analysis was performed using the model-based clustering method implemented in ADMIXTURE (Alexander et al., 2009) (version 1.3.0), which assigns individuals to K genetically homogeneous groups (i.e. K is the number of clusters) based on a subpopulation component value. The optimal clustering K value was estimated by setting from 1 to 15 to achieve different inferences, when the K value exhibits a lowest cross-validation error (CV error) value, each repeated 10 times. Then we used fastStructure (Raj et al., 2014) to confirm the optimal value of K.

Population structure was also inferred using the neighbor-joining tree method PHYLIP (version 3.698) (Retief 2000) and the tree layout was generated using the online tool iTOL (http://itol.embl.de). The population structure was further investigated by PCA (Reich et al., 2008) using GCTA (Yang et al., 2011) (Genome-wide Complex Trait Analysis, version 1.94.1), and the result from GCTA was plotted using an R script.

2.4 Measuring nucleotide diversity and differentiation

To assess genetic differentiation and diversity, we used the fixation index (F_{ST}) and nucleotide diversity (π) from the same 56,376 bi-allelic SNPs used in the population structure analysis. F_{ST} and π were calculated using VCFtools (version 0.1.15) (Danecek et al., 2011) with parameters --window-pi 500000 and --window-pi-step 50000 for the subgroups defined by ADMIXTURE. Additionally, we calculated linkage disequilibrium (LD) decay between each pair of SNPs using PopLDdecay (version 3.42) (Zhang et al., 2019).

2.5 Selective sweep and annotation of selected regions

To detect selective sweeps under artificial selection during domestication and improvement, we used the XP-CLR method (version 1.1.2) (Chen et al., 2010) with parameters set to a window size of 10 kb, a window step of 1 kb, and a maximum of 300 SNPs per window. Regions with scores in the top 5% were considered candidate selective regions.

Genes within these selected genomic regions were annotated to identify those undergoing strong selective sweeps. These genes were then subjected to enrichment analysis using the Gene Ontology (GO) (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000) databases, facilitated by the R package clusterProfiler (version 4.13.0) (Wu et al., 2021).

2.6 Genome-wide subpopulation ancestry and inter-subpopulation introgression inference

The procedure of subpopulation ancestry inference was as described previously, based on 3K-SNP and 3K-HAP datasets (Chen et al., 2020). The main steps were as follows: firstly, all sequenced cultivars were genotyped on the 3K-SNP sites with the GATK UnifiedGenotyper tool, adopting a "--output_mode EMIT_ALL_SITES" parameter. Then haplotypes were constructed by joining the SNP genotypes in each window. The NAF-scores of each haplotype were obtained by matching with 3K-HAP and the average value for each 100-kb window was taken. The subspecies or subpopulation ancestry of each window was inferred according to the NAF-score. For a sample that is assigned to a certain subpopulation, a window that has a different subspecies or subpopulation origin indicates a putative alien introgression.

2.7 Determination of 2-acetyl-1-phrroline in rice

Quantitative measurement of 2-acetyl-1-phrroline (2-AP) was achieved using the gas chromatography-tandem mass spectrometry (GC-MS) method (Peng et al., 2023). Firstly, a 250-g sample of rice was taken and pulverized through a standard sieve (25 mm mesh size). The pulverized sample was then transferred into a sample bottle and sealed with a sealing film at 4 °C for preservation.

For the pre-treatment of the sample, 1 g of the sample, accurately weighed to 0.01 g, was placed into a 10-mL centrifuge tube. To this, 1.5 mL of anhydrous ethanol was added, followed by immediate tightening of the tube cap. The mixture was thoroughly vortexed and then subjected to ultrasonication at 68 °C for 2 h in a water bath. After ultrasonication, the tube underwent centrifugation at 10,000 rpm for 10 min at 4 °C, with the resulting supernatant collected. This supernatant was further filtered through a microporous membrane (0.22

μm) before analysis.

Measurement was conducted under specific instrument reference conditions. Gas chromatography (GC) was performed using a capillary column with a stationary phase comprising 5% phenyl-95% methylpolysiloxane. The column temperature was programmed from an initial 45 °C, held for 1 min, then ramped at 8 °C/min to 100 °C, followed by a ramp at 50 °C/min to 250 °C, with a 1-min hold. Helium, with a purity of ≥99.999%, served as the carrier gas at a flow rate of 1.2 mL/min. The injector temperature was set at 250 °C, with a 1-mL injection volume using a splitless injection mode. Mass spectrometry (MS) was performed with an electron impact source at 70 eV and an ion source temperature of 250 °C. Interface temperature was maintained at 250 °C, with a solvent delay of 5.5 min. Acquisition was done via Multiple Reaction Monitoring (MRM). The retention time of 2-AP is about 6.094 min (Peng et al., 2023).

3 Results

3.1 Variation among the 376 temperate japonica rice elites

Sequencing of the genomes of 14 elite Hagengdao series rice cultivars and Hazhandao resulted in the generation of 1.35 billion paired-end reads, each about 150 base pairs long, with a sequencing depth averaging 31.88×. Furthermore, an additional 361 temperate *japonica* rice cultivars, selected from three separate studies conducted in HLJ province (Chen et al., 2023; Liu et al., 2021; Ye et al., 2022), were included, totaling a staggering 15.26 billion reads with an average sequencing depth of 15.55× (Table S1).

Alignment of all high-quality paired-end reads against the Nipponbare rice reference sequences (T2T-NIP, AGIS1.0) using BWA (Burrows-Wheeler Alignment, 0.7.17-r1188) revealed an impressive average mapping rate of 98.89%, ranging from 84.18% to 99.99%. These alignments covered 97.71% of the genome, with coverage ranging from 97.01% to 98.87%.

Subsequent analysis identified a total of 4.9 million SNPs and 0.98 million insertions and deletions (InDels) among the 376 elite cultivars, using the Genome Analysis Toolkit (GATK, version 4.0.5.1). Notably, most of these SNPs (59.19%) were located in intergenic regions, with only 40.81% residing within gene regions. Within the gene regions, 1.16% and 1.92% of the SNPs were found in the 5'- and 3'-untranslated regions (UTRs), respectively, while 20.37% were located in exons and 17.35% in introns (Table 1). Additionally, 70,583 InDels were located in exons and 207,885 in intro regions (Table S2).

Further analysis within coding regions revealed 589,531 missense variants, 363,547 synonymous variants, 35,288 splice region variants, 1,406 start-lost mutations, 32,779 stop-gained mutations, and 12,986 stop-lost mutations (Table S3).

Table 1 Number of SNPs on each chromosome in 376 Heilongjiang varieties

Chromosome	Intergenic	5’-UTR	3’-UTR	Exon	Intron	Total
Chr1	199735	4960	8938	68544	62572	344749
Chr2	249988	5683	9408	76412	75662	417153
Chr3	247019	6177	10576	66294	71560	401626
Chr4	245774	4755	5978	98143	74056	428706
Chr5	221450	4262	7231	74122	60515	367580
Chr6	269320	4500	8038	87624	75658	445140
Chr7	198882	3781	6989	68632	63503	341787
Chr8	263969	4401	8058	94173	79304	449905
Chr9	170951	2430	3926	59185	48877	285369
Chr10	295566	5829	8121	98007	77664	485187
Chr11	319502	6508	10705	126336	98007	561058
Chr12	226831	3819	6584	83822	65419	386475
Total	2908987	57105	94552	1001294	852797	4914735
Percent	59.19%	1.16%	1.92%	20.37%	17.35%	100.00%

UTR: Untranslated Regions

3.2 Population structure of the 376 elites

First, we performed genome-wide subpopulation ancestry inference using SNP markers derived from the 3K-RG project (Chen et al., 2020). As inferred by 3K-RG markers, all Hagengdao cultivars were inferred as being from temperate *japonica*, with *indica* introgression from 0.05% to 5.24% (Table S4).

Our study further corroborated this classification. Based on 56,376 SNPs from 376 elite cultivars, the population structure was elucidated using ADMIXTURE software with K values ranging from 1 to 15, representing the number of ancestral populations. The cross-validation (CV) error analysis showed a decreasing trend, making it challenging to pinpoint the optimal K value directly. Subsequently, we used fastStructure (version 20141213) software to determine the best K value, which was identified as K = 5. Consequently, the population was divided into five subgroups (Table S5). Contrary to previous studies, the cultivars released in 2010 and onwards were divided into two distinct categories in the Neighbor-Joining tree (Fig. 1a), designated HLJ-IV-1 and HLJ-IV-2, with all the Hagengdao series grouped within HLJ-IV-1. Compared with the previous classification of HLJ-IV as a single subgroup (only 43 samples in total), the HLJ-IV-1 and HLJ-IV-2 subgroups include more than 130 new cultivars, so they have higher reliability. Remarkably, more than half of the Longgeng series (63.4%) are grouped in HLJ-IV-2, and more than half of the Dongnong series (55.9%) in HLJ-IV-1. Furthermore, the breeding years within the subgroups did not completely align with the predefined classifications.

In HLJ-I, 38 out of 46 cultivars (82.6%) were from ~1980, including some from 2000-2010 like Kendao21 and Longgeng15. In HLJ-II, 25 out of 52 cultivars (48.1%) were from 1980-2000. HLJ-III had 53 cultivars (53.5% of the total 99), HLJ-IV-1 had 43 (44.8% of the total 96), and HLJ-IV-2 had 46 (55.4% of the total 83). PCA analysis indicated a somewhat weak population structure, with the top 10 principal components (PCs) explaining only 17.45% of the total variation (Fig. 1b). The PCA plot also clearly showed the separation between groups HLJ-IV-1 and HLJ-IV-2. The nucleotide diversity (π) of HLJ-IV-1 and HLJ-IV-2 was very similar, although the sample sizes of HLJ-IV-2 were smaller and exhibited higher values than other subgroups, suggesting that the cultivars developed in recent years have more diverse genetic sources (Fig. 1c). The PCA and population structure analysis revealed that HLJ-II and HLJ-III are very close and difficult to distinguish (Fig. 1e). Linkage disequilibrium (LD) decay rates, estimated as the physical distance at which LD dropped to half its maximum value, showed that HLJ-IV-1 decayed the fastest (Fig. 1d), suggesting that those cultivars had the highest recombination rate, likely due to their higher genetic diversity. We conducted a genome-wide scan for genetic differentiation (F_{ST}) in a pairwise manner among the five subgroups. Significant genetic differentiation was observed between each subgroup and HLJ-I, as well as between HLJ-IV-1 and HLJ-IV-2.

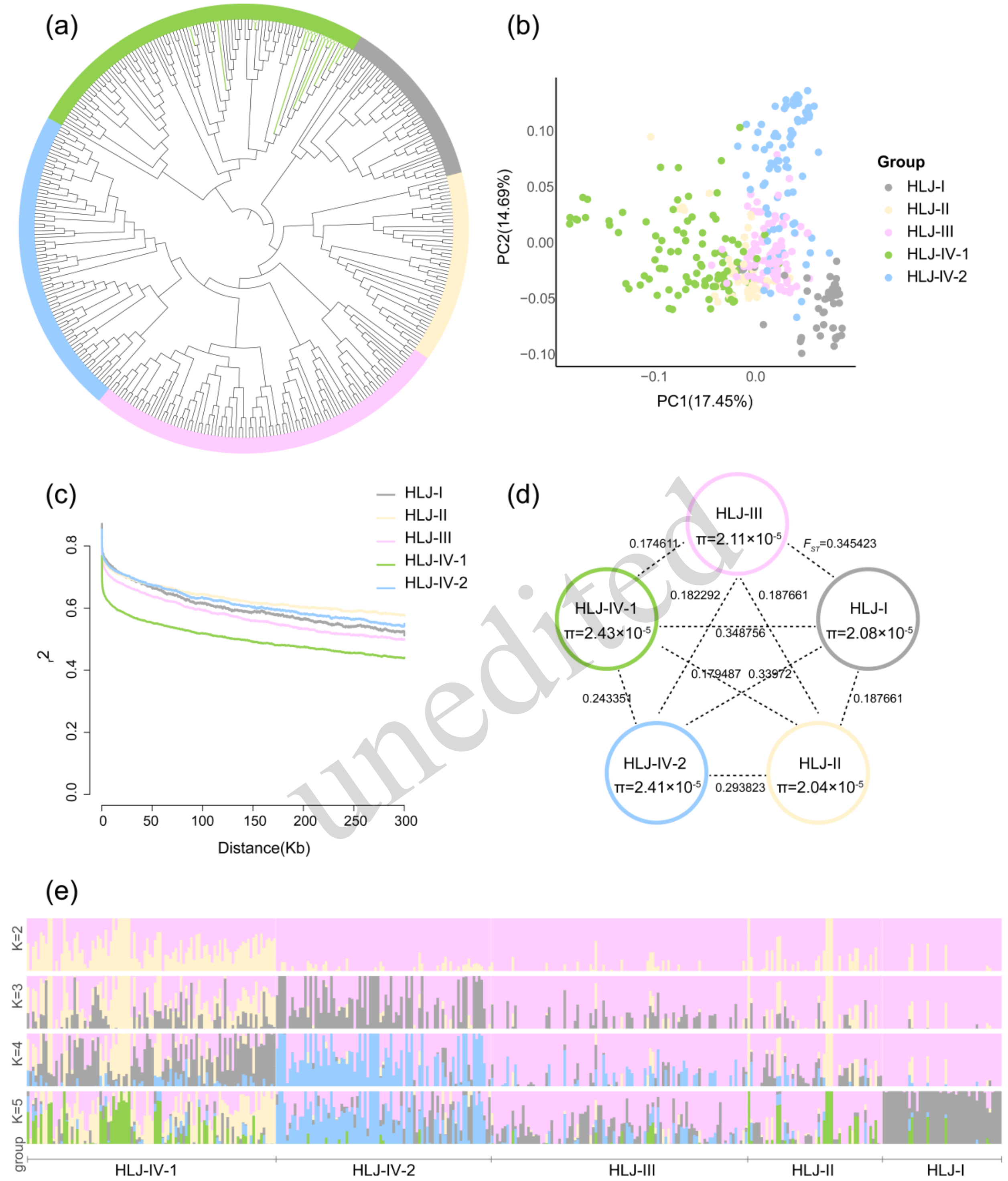

Fig. 1. Inferred population structure of HLJ rice cultivars. (a) Neighbor-join tree of samples from five inferred groups. The green lines represent the cultivars of the Hagengdao series. (b) PCA plot of all cultivars. (c) LD decay-distance analysis. (d) Genome-wide average π values in each subgroup and F_{ST} values between each pair of subgroups. (e) Subgroups inferred by using Admixture software of 376 cultivars using genome-wide SNP markers (*K*=2-5).

3.3 Different selection preferences of different subpopulations

Previous reports indicate that most traits of hybrid cultivars, such as heading date, source and sink organ traits, and grain quality, have changed significantly over time (Gu et al., 2023). We also noticed this trend in the cultivars from HLJ, particularly between HLJ-I and HLJ-IV, which span the longest time period. Therefore, we selected HLJ-I and HLJ-IV, along with HLJ-IV-1 and HLJ-IV-2, to compare the genes subjected to selection during the breeding process.

A selective sweep is an important natural selection pattern that fixes favorable mutation sites (Cutter and Payseur, 2013). To identify breeding targets, we screened signals of selective sweeps in subgenomes using XP-CLR scores. Each analysis revealed a diverse range of loci targeted between adjacent breeding periods in HLJ rice cultivars. Remarkably, the regions under selection comprised more than 10% of the entire genome. Among these, many commonly selected genes were associated with disease resistance. For instance, rice blast disease, caused by *Magnaporthe oryzae* (Ascomycota), is prevalent in about 80 countries and is considered one of the most devastating fungal diseases affecting rice (Valent, 2021). Genes such as *OsACBP5* (Narayanan et al., 2020) and *Os4CL5* (Gui et al., 2011), which enhance disease resistance via the jasmonic acid and lignin pathways respectively, were frequently selected (Fig. 2a). This indicates that selecting for broad-spectrum disease resistance has consistently been a key breeding objective. Furthermore, genes involved in grain filling and starch synthesis, like *GFR1* (Liu et al., 2019) and *Flo5* (Ryoo et al., 2007), were also under strong positive selection pressure (Fig. 2b). When examining genes specifically selected during different breeding periods, distinct patterns emerged. The putative selected genomic loci between HLJ-IV-1 and HLJ-IV-2 included genes related to tillering and plant architecture, such as *HTD2* (Liu et al., 2009) and *OsbZIP49* (Ding et al., 2021), reflecting efforts to optimize plant structure for higher yields and better adaptability. Loci associated with salt and cold tolerance were also selected, highlighting the need to develop cultivars capable of withstanding abiotic stresses specific to the regional growing conditions. In the putative selected genomic loci between HLJ-IV-1 and HLJ-I, some loci were associated with grain shape, such as *GW5* (Liu et al., 2017) and *GW6* (Shi et al., 2020).

The selected genetic regions are likely associated with specific agronomic traits, prompting us to perform KEGG pathway analysis. Comparing HLJ-IV-1 and HLJ-I, we found an enrichment of the carbohydrate metabolism pathway, which is crucial for developmental growth and yield-related traits (Wang et al., 2021). Comparing HLJ-IV-1 and HLJ-IV-2, we identified an enrichment of the glutathione metabolism and amino sugar and nucleotide sugar metabolism pathways, which are related to salt and cold tolerance, respectively (Yang et al., 2022). These findings further validate the functional roles of the identified genes, supporting their involvement in critical agronomic traits and stress responses.

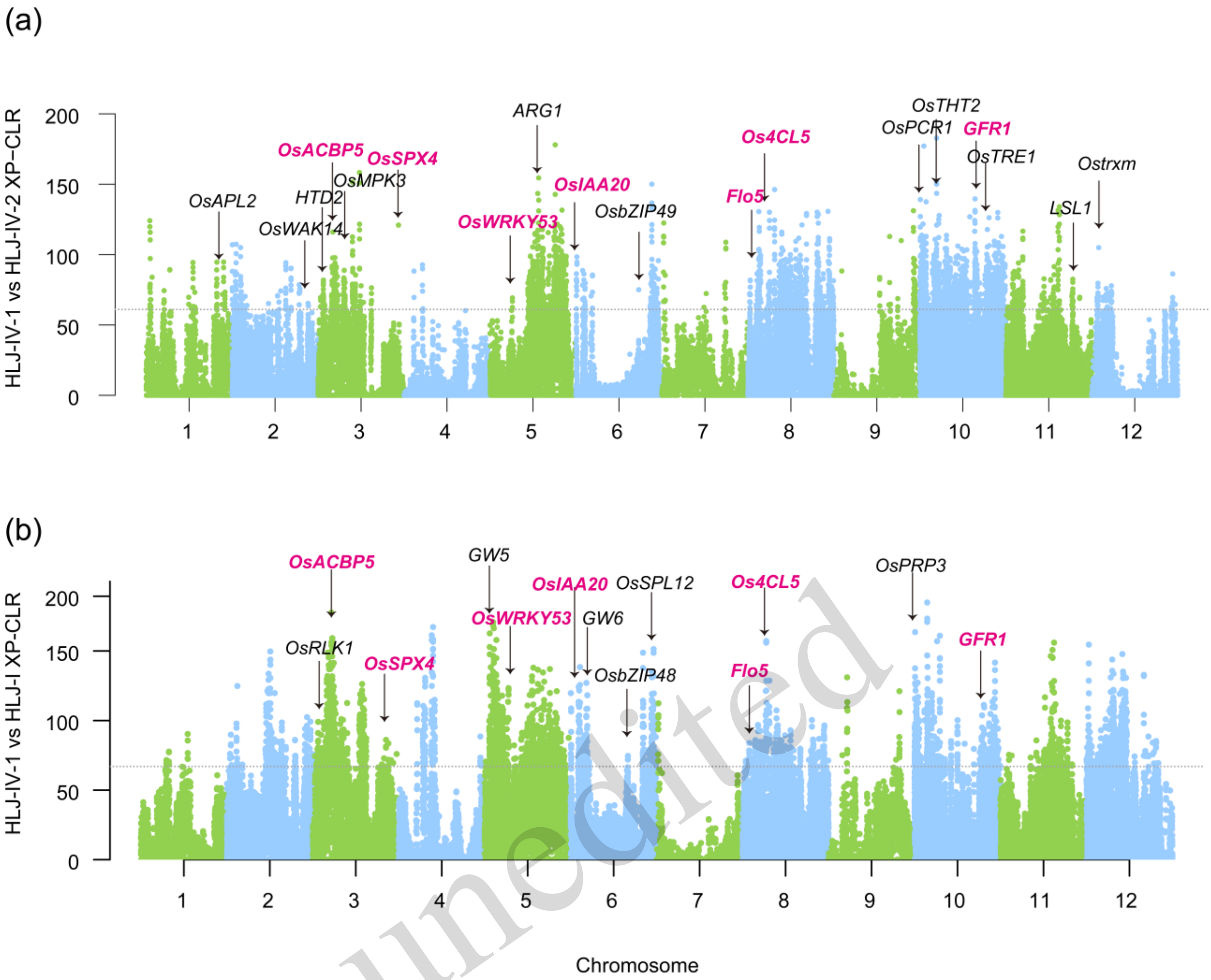

Fig. 2 Identification of putatively selected genomic regions based on the cross-population composite likelihood ratio (XP-CLR) scores in different stages of HLJ's rice cultivars. (a) putative selected genomic loci between HLJ-IV-1 and vs HLJ-IV-2. (b) putative selected genomic loci between HLJ-IV-1 and vs HLJ-I. The gene names in red indicate they are commonly selected.

3.4 Characteristics of the Hagengdao series cultivars

Building upon the insights gained from analyzing the 376 cultivars, we turn our focus to the Hagengdao series, which offers a unique opportunity to explore specific genetic traits and especially their contributions to aroma. Recently, the Harbin Academy of Agricultural Sciences developed a series of premium fragrant rice cultivars, including Hagengdao1, 2, 3, 4, 6, 7, 8, 9, 10, 11, 15, 16, 17, 18, and Hazhandao1. The pedigree charts of these cultivars (Fig. 3a) show that Wuyoudao 4, also known as DHX, serves as a crucial parental source. Notably, Wuyoudao 4 was derived from a variant of Wuyoudao 1 through pedigree selection. Similarly, Hagengdao1, 2, 3, and 7 were developed from field variants through selection processes. Consequently, these cultivars share several core characteristics, highlighting the institution's focus and achievements in breeding rice cultivars that have high yield, quality, and adaptability.

The phylogenetic tree constructed from Hagengdao cultivars, including Wuyoudao1 and Wuyoudao4, shows their genetic relationships and divergence. Most (11 out of 14) Hagengdao cultivars cluster with Wuyoudao4, while Ha1, Ha3, and Ha6 exhibit greater genetic distances from the other cultivars, particularly Ha1. This aligns with the results depicted in the pedigree diagram (Fig. 3b). Notably, Ha3 and Wuyoudao1 cluster closely together in the phylogenetic tree, reflecting their shared genetic background and close breeding relationship. Similarly, Ha4 and Ha16, whose parental sources include both WuyoudaoA and Wuyoudao4, also have close genetic ties. This consistency between pedigree and phylogenetic analyses underscores the robustness of the genomic analysis for determining the relationships among cultivars without pedigree information. These findings also provide insights into the genetic markers associated with desirable agronomic traits. However, the cultivars developed this year show relatively close genetic distances. Therefore, future breeding programs should aim to introduce superior alleles from other lines.

According to Ricebase (http://ricebase.org, accessed on June 1, 2024), the Hagengdao series outperforms control cultivars in regional trials, with a yield increase of 6.2-9.4%. Among these cultivars, Hagengdao15, with a plant height of 106.3 cm and a thousand-grain weight of 27.4 g, shows excellent tillering capacity and superior rice quality traits. Notably, Hagengdao7 is the only ultra-long grain cultivar in the series, though it suffers from significant lodging issues.

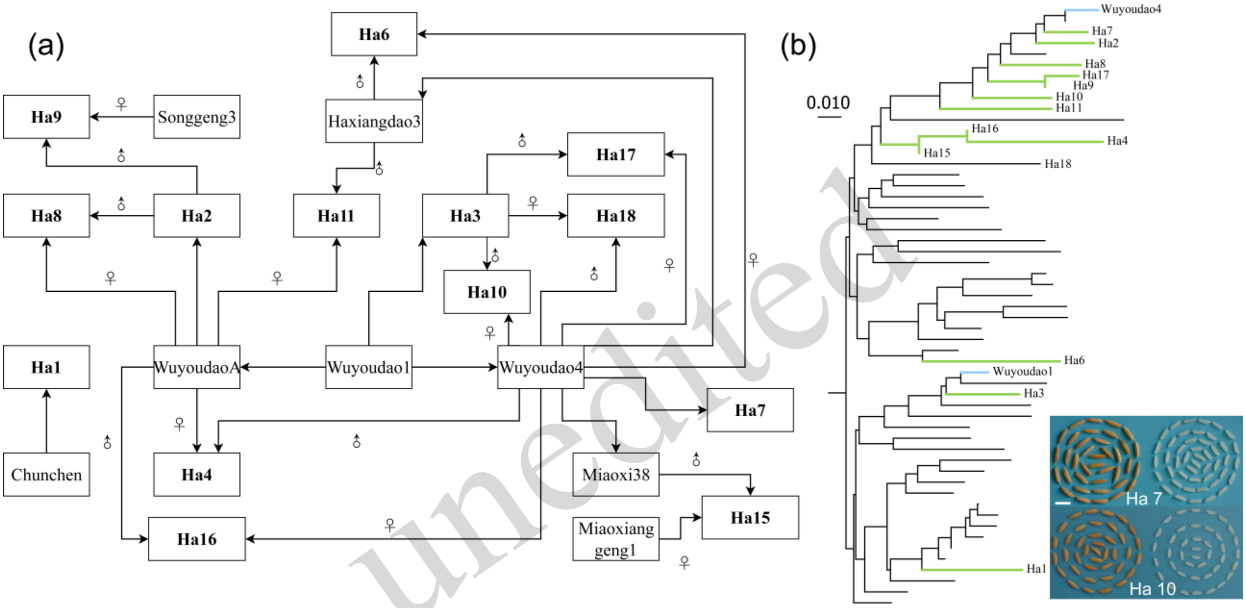

Fig. 3 Pedigree chart and important characteristics of the Hagengdao series cultivars. (a) Pedigree chart of the Hagengdao series cultivars, where "Ha" represents Hagengdao. The arrows indicate the direction of breeding crosses, while the male (♂) and female (♀) symbols denote the parental lines used in each cross. (b) Phylogenetic tree of Hagengdao and other derivatives. The tree illustrates the genetic distances and relationships among the cultivars, and the scale bar represents genetic divergence. The inset image shows Hagengdao7 has the characteristic of extra-long grains, which is not present in other cultivars of the Hagengdao series.

3.5 Genetic variation and aromatic compound analysis in Hagengdao

The breeding of fragrant rice in HLJ Province started relatively late, with the first fragrant *japonica* hybrid cultivar, Suigeng4, developed by the Suihua Branch of Heilongjiang Academy of Agricultural Sciences in 1999. Subsequently, a series of fragrant rice cultivars derived from Suigeng4 and Wuyoudao1 have been bred, leading to a rapid expansion in cultivation area, now exceeding 700,000 ha annually. Notably, the long-grain fragrant rice cultivar Suigeng4 and DHX have gained nationwide recognition for their exceptional quality.

The Hagengdao rice series, renowned for its popcorn-like aroma, has undergone extensive sensory evaluation and gas chromatography coupled with mass spectrometry (GC-MS) analyses. In addition to Hagengdao1 and 3, this characteristic aroma was detected across the series. The molecule 2-acetyl-1-pyrroline (2-AP) emerged as the primary aromatic compound, regulated by the genes betaine aldehyde dehydrogenase 2 (*OsBadh2*), alongside *OsBadh1*, *OsGly*, and *OsP5CS* (Prodhan and Qingyao 2020) (Fig. 4a). A recent study identified a natural 22-bp deletion in the coding region of the *OsODC* gene (Li et al., 2024). Furthermore, simultaneous knockout of both *OsBadh2* and *OsODC* genes can significantly enhance the 2-AP content in rice, with increases of up to 48% compared to cultivars lacking *OsBadh2* alone. This intricate genetic network governs the concentration of 2-AP, crucial for defining the aromatic profile of rice. Quantitative GC-MS analysis revealed that Hagengdao7 had the highest 2-AP content, reaching 0.0515 mg/kg, while DHX and Hagengdao4 had concentrations of 0.03586 and 0.02704 mg/kg, respectively (Fig. 4b). Remarkably, Hagengdao1 and 3 showed nearly undetectable levels of 2-AP (Table S6). The functional enzyme encoded by *OsBadh1/2* facilitates the conversion of GAB-ald to GABA (Sakthivel et al., 2009), a pivotal step in inhibiting 2-AP synthesis in non-aromatic rice cultivars. Dysfunction in the badh2 enzyme disrupts this conversion, resulting in GAB-ald accumulation and subsequent 2-AP production.

A key revelation in the genetic makeup of most Hagengdao cultivars was the identification of specific mutations within the *OsBadh2* (*AGIS_Os03g038200*) and *OsODC* (*AGIS_Os08g030110*) genes. The *OsBadh2* gene exhibits single nucleotide mutations (A-T) and an 8-bp deletion at positions 20,519,996, 20,519,998, and 20,520,003 in the 7th exon, while *OsODC* shows a continuous 22-bp deletion starting at position 17,101,416 (Fig. 4d). Analysis of 376 Heilongjiang cultivars revealed four haplotypes formed by these genes. Hap1, which is identical to the reference genome, constitutes the largest proportion (80.85%), whereas Hap2, with mutations in both genes, represents the smallest proportion (2.13%). Hap3 and Hap4 correspond to cultivars with deletions in either *OsBadh2* or *OsODC*, respectively. The intermediate circular diagram categorizes Hap1-4 among 14 Hagengdao cultivars, with six cultivars having mutations in both genes, accounting for 75% of Hap1, aligning well with the observed 2-AP content. Despite the deletion of *OsODC* in Ha1, the absence of fragrance persists. This is because *Osbadh2* remains intact, further supporting the role of *OsODC* as a key enhancer of aroma. However, its influence is largely negligible in the presence of a functional *Osbadh2*. This supports the significant contribution of these mutations to the fragrant rice resources in Heilongjiang Province (Fig. 4c). The remaining two Hap1 cultivars are DHX and Longxiangdao2. These genetic alterations play a pivotal role in imparting the distinctive aroma to these rice cultivars. Further visualization of these gene variations using IGV (Integrative Genomics Viewer) has corroborated our findings (Fig. 4e).

This study confirmed the genetic determinants of rice aroma, particularly in Hagengdao7, highlighting its potential as a focal point in breeding efforts aimed at enhancing sensory quality. It lays the groundwork for targeted gene manipulation to develop rice cultivars with enhanced flavor profiles.


(a)

Gene	MSU_ID	T2T_ID	Reference
***BADH* gene**			
OsBadh2	*LOC_Os08g32870*	*AGIS_Os08g030110*	Lorieux *et al.*, 1996
OsBadh1	*LOC_Os04g39020*	*AGIS_Os04g033860*	Amarawathi *et al.*, 2008
***P5CS* gene**			
OsP5CS1	*LOC_Os05g38150*	*AGIS_Os05g033150*	Huang *et al.*, 2008
OsP5CS2	*LOC_Os01g62900*	*AGIS_Os01g055180*	Huang *et al.*, 2008
***GLY* gene**			
OsGlyI	*LOC_Os05g22970*	*AGIS_Os05g019380*	Talukdar *et al.*, 2017
OsGlyII	*LOC_Os03g21460*	*AGIS_Os03g019100*	Pachauri *et al.*, 2014
OsGlyIII	*LOC_Os01g47690*	*AGIS_Os01g040670*	Pachauri *et al.*, 2014
***ODC* gene**			
OsODC	*LOC_Os02g28110*	*AGIS_Os02g024830*	Li *et al.*, 2024

Fig. 4 Number of aromatic cultivars, aroma genes, and their mutation types in the Hagengdao and Heilongjiang series cultivars. (a) Description of potential genes associated with aroma in rice, including *BADH*, *P5CS*, *GLY* and *DOC*. (b) Content of 2-acetyl-1-pyrroline (2-AP). Different lowercase letters in the graphs indicate significant differences ($p<0.05$). (c) Four haplotypes formed by *Badh2* and *DOC* in 376 Heilongjiang cultivars. Hap1, identical to the reference genome, includes the largest proportion (80.85%), while Hap2, with mutations in both genes, represents the smallest proportion (2.13%) of the cultivars. The intermediate circular diagram categorizes Hap1-4 for 14 cultivars of Hagengdao. (d) Mutation types of aroma genes *OsDOC* and *Osbadh2*. Among the Hagengdao series, only Ha3 has no mutations in either gene. (e) IGV (Integrative Genomics Viewer) visualization of variations in *OsDOC* and *Osbadh2*.

3.6 Variation of other important function-defined genes

An important task for long-term breeding in HLJ is to thoroughly analyze the superior alleles located in important loci, because such insights can help significantly enhance the development of improved rice cultivars with desirable traits. As a superior germplasm resource, the cultivar Hagengdao7 exhibits an exceptional trait among *japonica* rice cultivars, boasting an extraordinary grain length-to-width ratio of 3.2:1, in contrast to other Hagengdao rice cultivars which typically have long grains with a ratio ranging from 2.6:1 to 2.8:1. Further elucidating its genetic basis, research has identified the rice PPKL family gene *OsGL3.1* (*AGIS_Os03g038200*) as a key regulator of seed size and yield (Long et al., 2023). Among 376 cultivars from Heilongjiang Province, only Hagengdao7 harbors a unique exon-level SNP at position 25111196 of the 11th exon, where a cytosine (C) is replaced by a thymine (T), introducing a premature stop codon that truncates translation, possibly leading to the remarkable extra-long grain trait observed in Hagengdao (Fig. 5a).

In the 1960s, the semi-dwarf gene *sd1* (Terao and Hirose, 2015) (*AGIS_Os01g058220*), coding for a gibberellin synthase, played a pivotal role in the "Green Revolution" of rice cultivation (Khush, 2001). *Indica* rice harbors null-function alleles (*sd1-d*, *sd1-AJNT*, *sd1-9311*, *sd1-bm*), whereas *japonica* rice carries weak-function alleles (*sd1-j*, *sd1-ZYQ8*, *sd1-c*, *sd1-r*). Throughout the course of evolution and human-mediated domestication, a weak-functioning allelic variant, *SD1-EQ* (*Sd1*^{Jap}), has been predominantly preserved in *japonica* rice. Among 376 series cultivars, those carrying the *sd1-r* allele in Heilongjiang *japonica* rice include eight Dongnong series cultivars (Dongnong419, Dongnong423, Dongnong427, Dongnong9008, Dongnong71-51, Dongnong6212, Dongnong4201, Dongnong426), Longdao10, and Mudanjiang21. A mutation at position 39015397 in the third exon, changing a G to a C, results in an amino acid alteration from Asp to His at position 348 (Fig. 5b). According to the MBKbase-Rice database (Peng et al., 2020) (www.mbkbase.org/rice, accessed 1 June 2024), these cultivars show a height reduction from an unmutated average of (102±13.3) cm to (91.7±7.1) cm, approximating a 10.1% dwarfism. Notably, Dongnong427 and Longdao21, derived from Dongnong423, exhibit superior lodging resistance in practical rice production. However, the lodging resistance of Hagengdao series cultivars, particularly Hagengdao7 and 10, is relatively weak. The Xiushui cultivar developed in Zhejiang Province, with its *sd1-j* allele, provides crucial genetic resources for dwarfing Hagengdao cultivars. In the breeding of semi-dwarf *japonica* rice, the *sd1-d* allele from *indica* rice was introduced into *japonica* rice DHX. Following the challenges posed by the typhoons of 2020, selection processes led to the development of dwarf and semi-dwarf lines carrying the *sd1-d* gene, named 1279 and 1280, respectively. These lines are expected to serve as valuable intermediates for future genetic research and breeding efforts. This work facilitates parental selection and marker-assisted breeding, laying a foundation for the identification of favorable genes co-selected with the *sd1* allele in Hagengdao parental lines.

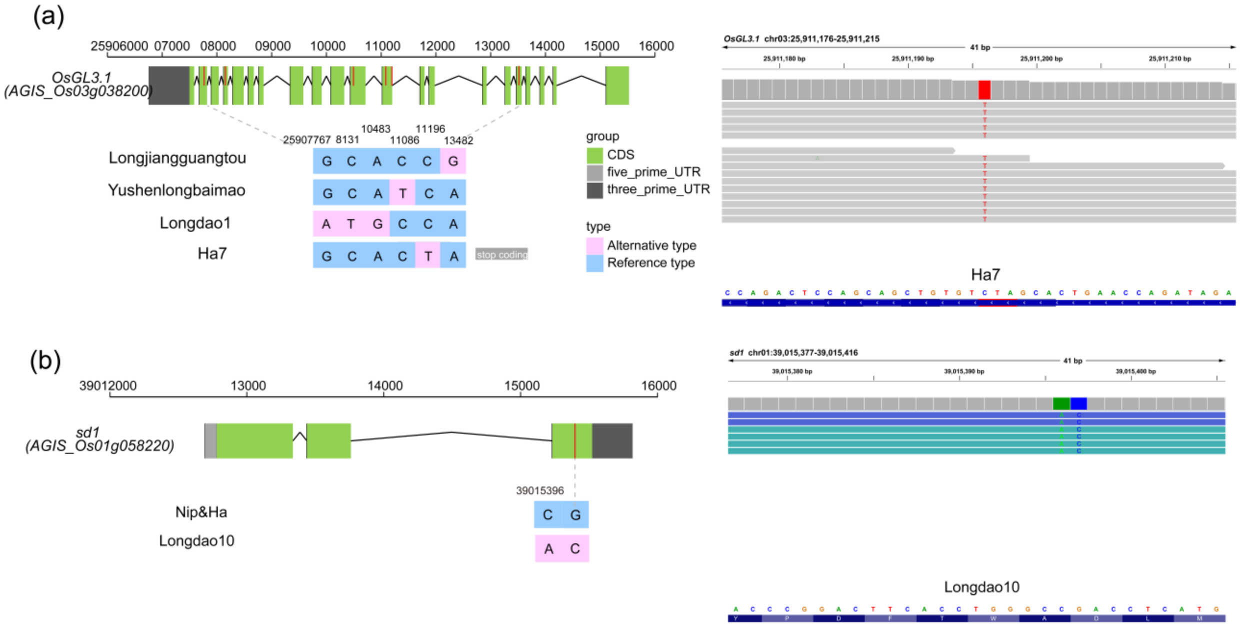

Fig. 5 Types of variation in important genes in the Hagengdao and Heilongjiang series cultivars including IGV (Integrative Genomics Viewer) visualization of variations in these genes. (a) Single-nucleotide polymorphisms of the grain length-related gene *OsGL3.1* in the Heilongjiang series cultivars. There are four haplotypes, with Hagengdao7 having a unique variant that leads to a premature stop codon in the exon of this gene. (b) Mutation types of the lodging resistance-related gene *sd1*. Similar to the reference genome, the Hagengdao series and *sd1-r* genotype are absent. Other Heilongjiang series cultivars, such as Longdao10, exhibit strong lodging resistance, attributed to the presence of *sd1-r*.

4 Discussion

The sequencing of the complete genome of Nipponbare, a significant rice cultivar, has provided a crucial benchmark for the genetic study of rice, enhancing our understanding of the relationship between specific genes and rice traits. Enabled by second-generation sequencing technology, the in-depth analysis of multiple rice genomes has identified numerous genetic variants. These variants offer valuable resources for improving specific traits in breeding, thus advancing the potential for generating superior recombinant genotypes. This study's decoding of 376 elite *japonica* rice genomes offers breeders comprehensive insights into both broad and detailed aspects of parental line genomes, including the effects of genomic variation, SNPs and InDels on gene functions, and allelic differences at functionally defined genes. This extensive genomic information significantly enhances parental selection and cross design, increasing the likelihood of generating superior recombinant genotypes aligned with breeding objectives from segregating populations.

Moreover, the analysis of the population structure of these 376 cultivars identified distinct subgroups among the elite cultivars, elucidating their genetic diversity and evolutionary relationships. Previous research classified the HLJ cultivars based on their year of release into four historical groups: HLJ-I (~1980), HLJ-II (1980-2000), HLJ-III (2000-2010), and HLJ-IV (2010~) (Chen et al., 2023). Our study corroborates this classification. Notably, by adding over 130 cultivars to the previous HLJ-IV subgroup, we increased its reliability, categorizing HLJ-IV cultivars into two distinct subgroups, HLJ-IV-1 and HLJ-IV-2, with most Hagengdao series in HLJ-IV-1. This detailed understanding of the population structure enables more strategic selection of parental lines, fostering the development of high-yielding, resilient rice cultivars. Additionally, the selection patterns observed in HLJ rice breeding support a multifaceted approach to enhancing both biotic and abiotic stress resistance, improving yield traits, and refining plant architecture to meet challenges posed by varying environmental conditions and disease pressures. The distinction between HLJ-II and HLJ-III in the PCA results appears somewhat unclear. This is to be expected, as differences among rice cultivars in HLJ are less pronounced than those across the broader northeast region of China. HLJ-II comprises 25 middle, 23 new, and 4 very new samples, while HLJ-III includes 8 early, 20 middle, 53 new, 11 very new, and 7 unknown samples. This distribution suggests that HLJ-III has a higher representation of newer cultivars, particularly in the 'new' category, which contributes to the observed overlap. Notably, over 70% of HLJ cultivars prior to the 1980s clustered within the Ishikari-Shiroge group, and both HLJ-II and HLJ-III incorporate Japanese backbone cultivars introduced during that decade, complicating their differentiation. Since 2000, the average content of *indica* introgressions in cultivars has increased significantly, alongside the breeding of stable, cold-resistant, and lodging-resistant cultivars such as Longgeng31, Daohuaxiang2, and Zhongkefa5 from 2013 onward (Chen et al., 2023), which have enhanced the differentiation of the HLJ-IV subgroup.

The rice cultivars developed by the Harbin Academy of Agricultural Sciences achieve a balance between yield and quality while incorporating resistance and adaptability. This balance enables these cultivars to meet the cultivation requirements of diverse ecological zones. The institution's profound expertise and foresight in enhancing rice quality and adaptability are demonstrated through their meticulous consideration of these traits. These rice series, particularly Hagengdao7, have shown significant breeding potential with their unique aromatic profiles and elongated grain size, but have poor lodging resistance. Future breeding strategies will focus on enhancing aromatic qualities, optimizing grain shape, improving lodging resistance, and expanding genetic diversity to increase the adaptability and stability of rice.

The impact of gene mutation on the diversity of aroma compounds remains a subject of considerable interest. Currently, among all the rice cultivars in HLJ, only the *Badh2* and *ODC* aroma gene mutations have been detected. No mutations have been found in the *Badh1*, *P5CS*, and *GLY* genes, which are also associated with the synthesis pathway of 2-AP. Particularly, for the homologous gene *Badh1*, it remains to be determined whether simultaneous mutations in both *Badh1* and *Badh2* using targeted gene editing techniques like CRISPR/Cas9 can further enhance fragrance. Thus, the limited genetic diversity of fragrant rice in HLJ reinforces the need to expand the genetic resources of fragrant rice to enrich its genetic foundation. Comparative analysis of fragrance between Hagengdao2 and DHX revealed minimal differences, while many other cultivars exhibited degradation in aroma compounds. Notably, inconsistencies were observed between the measured data and sensory evaluations, and different individuals had variable sensory assessments. This suggests that substances other than 2-AP may also play a role in the aroma of fragrant rice. Regarding the interplay between *ODC* and *Badh2*, the measured 2-AP content generally aligned with the haplotype analysis results, though there were exceptions. For instance, Ha4, which has a deletion in only *OsBadh2*, exhibits significantly higher 2-AP levels than Ha6, in which both genes were knocked out. This suggests that while these genes play critical roles in the pathway, there may be other undiscovered genes that also regulate 2-AP synthesis. Genome editing has become a powerful breeding technique, but it relies on the (full) understanding of the genetics and genomics of traits of interest. The finding of the present study thus is of great importance to the improvement of rice cultivars in HLJ Province. For instance, we discovered that Ha4 had a high 2-AP content but its *OsODC* gene was not knocked out, hence we could potentially produce superior germplasm with even higher 2-AP content than that found in Ha7 by knocking out the *OsODC* gene by genomic editing.

Meanwhile, optimizing grain shape requires considering factors beyond length, exploring the function of *OsGL3.1* and other relevant genes to identify additional genetic targets. Finally, enhancing lodging resistance hinges on a comprehensive understanding of the role of the *sd1* gene and its allelic variants, coupled with adaptability studies to cultivate high-quality, resilient cultivars. However, recent climatic shifts in the northeast, widespread adoption of direct-seeding methods, and ambitious high-yield breeding strategies have intensified the demand for lodging resistance in *japonica* rice cultivars, including the Harbin *japonica* series and other cultivars from HLJ Province. The Dongnong and Xiushui series, derived from the foundational *japonica* parents Qiuguang and Ce21, respectively, harbor the *sd1-r* and *sd1-j* alleles, showcasing their effective application in contemporary *japonica* rice breeding. Expanding genetic diversity through the introduction of superior aromatic rice resources from domestic and international origins, alongside the use of efficient molecular markers and genome selection techniques, will expedite the breeding process.

Data availability statement

All raw reads generated for rice accessions used in this study have been deposited in the National Genomics Data Center with BioProject PRJCA024912 and GSA accession CRA025120, which is publicly accessible at https://ngdc.cncb.ac.cn/.

Acknowledgments

This study was partially financially supported by Central Leading Local Science and Technology Development Project (ZY2022A-HRB-02), Heilongjiang Postdoctoral Science Foundation (LBH-Z22245), Harbin Science and Technology Research (College Cooperation) Project (GJ2021TZ002007), Basic Scientific Research in Colleges and Universities in Heilongjiang Province (2020-KYYWF-1028).

Author contributions

Qingtao YU and Qingyao SHU designed and led this project. Yuhan ZHOU and Sanling WU re-sequenced the genome. Yuhan ZHOU, Naixin LIU, Jiaqi YANG, Baicui CHEN, Ziqi ZHOU and Chengxin LI and Fanshan BU analyzed the data. Yuhan ZHOU and Qingyao SHU wrote the draft manuscript. Yuhan ZHOU, Qingtao YU and Qingyao SHU discussed and revised the draft. All authors have read and agreed to the published version of the manuscript.

Compliance with ethics guidelines

Yuhan Zhou, Naixin Liu, Jiaqi Yang, Baicui Chen, Chengxin Li, Fanshan Bu, Sanling Wu, Ziqi Zhou, Qingtao Yu and

Qingyao Shu declare that they have no conflict of interest. This article does not contain any studies with human or animal subjects performed by any of the authors.

References

Alexander DH, Novembre J, Lange K, 2009. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res*, 19(9):1655-1664.
https://doi.org/10.1101/gr.094052.109

Ashburner M, Ball CA, Blake JA, et al., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, 25(1):25-29.
https://doi.org/10.1038/75556

Bolger AM, Lohse M, Usadel B, 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, 30(15):2114-2120.
https://doi.org/10.1093/bioinformatics/btu170

Chen H, Patterson N, Reich D, 2010. Population differentiation as a test for selective sweeps. *Genome Res*, 20(3):393-402.
https://doi.org/10.1101/gr.100545.109

Chen Z, Bu QY, Liu GF, et al., 2023. Genomic decoding of breeding history to guide breeding-by-design in rice. *Natl Sci Rev*, 10(5):nwad029.
https://doi.org/10.1093/nsr/nwad029

Chen Z, Li XX, Lu HW, et al, 2020. Genomic atlases of introgression and differentiation reveal breeding footprints in Chinese cultivated rice. *J Genet Genomics*, 47(10):637-649.
https://doi.org/10.1016/j.jgg.2020.10.006

Cingolani P, Platts A, Wang le L, et al., 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)*, 6(2):80-92.
https://doi.org/10.4161/fly.19695

Cutter AD, Payseur BA, 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet*, 14(4):262-274.
https://doi.org/10.1038/nrg3425

Danecek P, Auton A, Abecasis G, et al., 2011. The variant call format and VCFtools. *Bioinformatics*, 27(15):2156-2158.
https://doi.org/10.1093/bioinformatics/btr330

DePristo MA, Banks E, Poplin R, et al., 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet*, 43(5):491-498.
https://doi.org/10.1038/ng.806

Ding CH, Lin XH, Zuo Y, et al., 2021. Transcription factor OsbZIP49 controls tiller angle and plant architecture through the induction of indole-3-acetic acid-amido synthetases in rice. *Plant J*, 108(5):1346-1364.
https://doi.org/10.1111/tpj.15515

Gu ZL, Gong JY, Zhu Z, et al., 2023. Structure and function of rice hybrid genomes reveal genetic basis and optimal performance of heterosis. *Nat Genet*, 55(10):1745-1756.
https://doi.org/10.1038/s41588-023-01495-8

Gui JS, Shen JH, Li LG, 2011. Functional characterization of evolutionarily divergent 4-Coumarate:coenzyme a ligases in rice. *Plant Physiology*, 157(2):574-586.
https://doi.org/10.1104/pp.111.178301

Ikeda M, Miura K, Aya K, et al., 2013. Genes offering the potential for designing yield-related traits in rice. *Curr Opin Plant Biol*, 16(2):213-220.
https://doi.org/10.1016/j.pbi.2013.02.002

Jie Y, Shi TY, Zhang ZJ, et al., 2021. Identification of key volatiles differentiating aromatic rice cultivars using an untargeted metabolomics approach. *Metabolites*, 11(8):528.
https://doi.org/10.3390/metabo11080528

Kanehisa M, Goto S, 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27-30.
https://doi.org/10.1093/nar/28.1.27

Khush GS, 2001. Green revolution: the way forward. *Nat Rev Genet*, 2(10):815-822.
https://doi.org/10.1038/35093585

Li H, Durbin R, 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14):1754-1760.
https://doi.org/10.1093/bioinformatics/btp324

Li H, Durbin R, 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics*, 26(5):589-595.
https://doi.org/10.1093/bioinformatics/btp698

Li H, Handsaker B, Wysoker A, et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*,

25(16):2078-2079.
https://doi.org/10.1093/bioinformatics/btp352

Li Y, Zhang WT, Li MY, et al., 2024. Discovery of *OsODC* as a key enhancer of aroma and development of highly fragrant rice. Plant Commun.
https://doi.org/10.1016/j.xplc.2024.101141.

Liu CX, Peng P, Li WG, et al., 2021. Deciphering variation of 239 elite japonica rice genomes for whole genome sequences-enabled breeding. *Genomics*, 113(5):3083-3091.
https://doi.org/10.1016/j.ygeno.2021.07.002

Liu EB, Zeng SY, Zhu SS, et al., 2019. Favorable alleles of GRAIN-FILLING RATE1 increase the grain-filling rate and yield of rice. *Plant Physiol*, 181(3):1207-1222.
https://doi.org/10.1104/pp.19.00413

Liu JF, Chen J, Zheng XM, et al., 2017. GW5 acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. *Nat Plants*, 3:17043.
https://doi.org/10.1038/nplants.2017.43

Liu WZ, Wu C, Fu YP, et al., 2009. Identification and characterization of HTD2: a novel gene negatively regulating tiller bud outgrowth in rice. *Planta*, 230(4):649-658.
https://doi.org/10.1007/s00425-009-0975-6

Long Y, Wang C, Liu C, et al., 2023. Molecular mechanisms controlling grain size and weight and their biotechnological breeding applications in maize and other cereal crops. *J Adv Res*, 62(7):27-46.
https://doi.org/10.1016/j.jare.2023.09.016

Panthapulakkal Narayanan S, Lung SC, Liao P, et al., 2020. The overexpression of OsACBP5 protects transgenic rice against necrotrophic, hemibiotrophic and biotrophic pathogens. *Sci Rep*, 10(1):14918.
https://doi.org/10.1038/s41598-020-71851-9

Peng H, Wang K, Chen Z, et al., 2020. MBKbase for rice: an integrated omics knowledgebase for molecular breeding in rice. *Nucleic Acids Res*, 48(D1):D1085-D1092.
https://doi.org/10.1093/nar/gkz921

Peng JF, Zhu Y, Lin F, et al., 2023. Direct determination of 2-Acetyl-1-Pyrroline in rice by ultrasound-assisted solvent extraction coupled with ultra-performance liquid chromatography-tandem mass spectrometry. *Food Analytical Methods*, 16(5):900-908.
https://doi.org/10.1007/s12161-023-02478-5

Prodhan ZH, Shu QY, 2020. Rice aroma: A natural gift comes with price and the way forward. *Rice Science*, 27(2):86-100. https://doi.org/10.1016/j.rsci.2020.01.001

Purcell S, Neale B, Todd-Brown K, et al., 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559-575.
https://doi.org/10.1086/519795

Raj A, Stephens M, Pritchard JK, 2014. fastSTRUCTURE: variational inference of population structure in large SNP data sets. *Genetics*, 197(2):573-589.
https://doi.org/10.1534/genetics.114.164350

Reich D, Price AL, Patterson N, 2008. Principal component analysis of genetic data. *Nat Genet*, 40(5):491-492. https://doi.org/10.1038/ng0508-491

Retief JD, 2000. Phylogenetic analysis using PHYLIP. *Methods Mol Biol*, 132:243-258.
https://doi.org/10.1385/1-59259-192-2:243

Ryoo N, Yu C, Park CS, et al., 2007. Knockout of a starch synthase gene *OsSSIIIa/Flo5* causes white-core floury endosperm in rice (*Oryza sativa* L.). *Plant Cell Rep*, 26(7):1083-1095.
https://doi.org/10.1007/s00299-007-0309-8

Sakthivel K, Sundaram RM, Shobha Rani N, et al., 2009. Genetic and molecular basis of fragrance in rice. *Biotechnol Adv*, 27(4):468-473.
https://doi.org/10.1016/j.biotechadv.2009.04.001

Shang LG, He WC, Wang TY, et al., 2023. A complete assembly of the rice Nipponbare reference genome. *Mol Plant*, 16(8):1232-1236.
https://doi.org/10.1016/j.molp.2023.08.003

Shi CL, Dong NQ, Guo T, et al., 2020. A quantitative trait locus GW6 controls rice grain size and yield through the gibberellin pathway. *Plant J*, 103(3):1174-1188.
https://doi.org/10.1111/tpj.14793

Terao T, Hirose T, 2015. Control of grain protein contents through SEMIDWARF1 mutant alleles: *sd1* increases the grain protein content in Dee-geo-woo-gen but not in Reimei. *Mol Genet Genomics*, 290(3):939-954.

https://doi.org/10.1007/s00438-014-0965-7

Valent B, 2021. The impact of blast disease: past, present, and future. *Methods Mol Biol*, 2356:1-18. https://doi.org/10.1007/978-1-0716-1613-0_1

Supplementary information

Tables S1-S6